THE ROLE OF ENRICHMENT IN DECREASING DIVERSITY OF MYCOBACTERIOPHAGE GENOME DATABASES

by
Vanessa Gonzalez

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science

Baltimore, Maryland
April 2019

Abstract

While there is great genetic diversity among phages, a large proportion of mycobacteriopphages fall into only a few clusters. Does the observed distribution of members in clusters actually reflect what is present in nature or does the enrichment procedure cause a skew in diversity? We hypothesize that the enrichment procedure promotes the replication of phages belonging to only a few clusters, thus decreasing the diversity of phages identified from a sample. Using nanopore sequencing to conduct a metagenomics analysis of soil samples, a decrease in the number of clusters present in enriched samples, compared to unenriched samples, was observed. The data supports the hypothesis and demonstrates that enrichment promotes the growth of only a select few clusters and subclusters, making it more likely to isolate phages of these clusters and subclusters. With the growing potential and prevalence of phage applications, it is important to expand our knowledge on their diversity. Conducting studies on phage diversity not only leads to a larger array of phages to use for applications, but also gives us more information on how phages interact with their environment.

Advisors: Dr. Joel Schildbach and Dr. Robert Horner

Preface

This thesis is the product of my Master's research initiated in August 2018 and completed in April 2019. The research was done in the Johns Hopkins University in Baltimore, Maryland.

I was first introduced to bacteriophages and their isolation, characterization, and genomics by Dr. Joel Schildbach and Dr. Emily Fisher in their Phage Hunting Lab. After learning about bacteriophages in class, I became interested in further studying bacteriophages as a member of Dr. Schildbach's lab. Dr. Schildbach expressed his interest in exploring the enrichment procedure and its potential effects on phage diversity in soil samples. I too became intrigued by this idea and decided to pursue this as my project.

I would like to thank Dr. Schildbach for his continuous mentorship, guidance, and support throughout this study. Additionally, I would like to thank my lab members, particularly Russell Hughes, Amy Nguyen, and Shanna Leventhal, for their help and support. I would also like to thank Dr. Winston Timp and his lab's members, particularly Norah Sadowski and Yunfan Fan, for their collaboration and expertise in nanopore sequencing. I would also like thank Dr. Kathryn Tifft Oshinnaiye for her continuous guidance and support throughout this study. Finally, I would like to thank my dear family and friends, who have supported me throughout my education and research.

Table of Contents

List of Figures

Chapter 1: Phage diversity and applications

Introduction

The following review chapter discusses various topics regarding bacteriophages. Bacteriophages, also known as phages, are a class of viruses that infect and kill bacteria. There is an incredibly large amount of diversity among phages present in the world, although it remains relatively unstudied. Despite our limited knowledge on phages, a variety of phage applications have been developed. Therefore, studying phages will expand our knowledge and further our applications.

Phage structure and reproduction

There are many different types of phages, each containing variations in structure. Tailed phages consist of three basic components: the capsid or head, the tail, and the genetic material (Fig. 1). The capsid contains the genetic material, which is typically double-stranded DNA.

Phages infect bacteria by injecting their DNA to initiate their reproduction within the bacterium. Specifically, the phage will bind to surface receptors of a bacterium. The binding of a phage to a bacterium is very specific, with a phage only being able to bind if the bacterium contains receptors the phage can bind to. When bound to a bacterium, the phage injects its genetic material. The bacterial machinery will begin to replicate and express the phage genome, resulting in the production of phage proteins. Phage proteins then assemble into complete phage particles. Additional phage proteins function to degrade the cell membrane and wall of the bacterium, causing an influx of liquids and lysis [1]. The new phages are released and begin the lytic cycle anew.

Many phages, particularly temperate phages, are also able to undergo a lysogenic cycle [2]. The lysogenic cycle also involves the injection of genetic material into a bacterial host, but, instead lysis, it results in the integration of the phage's genome into the bacterial genome. The phage genome will be replicated along with the host genome until the phage genome excises from the bacterial genome and initiates the lytic cycle.

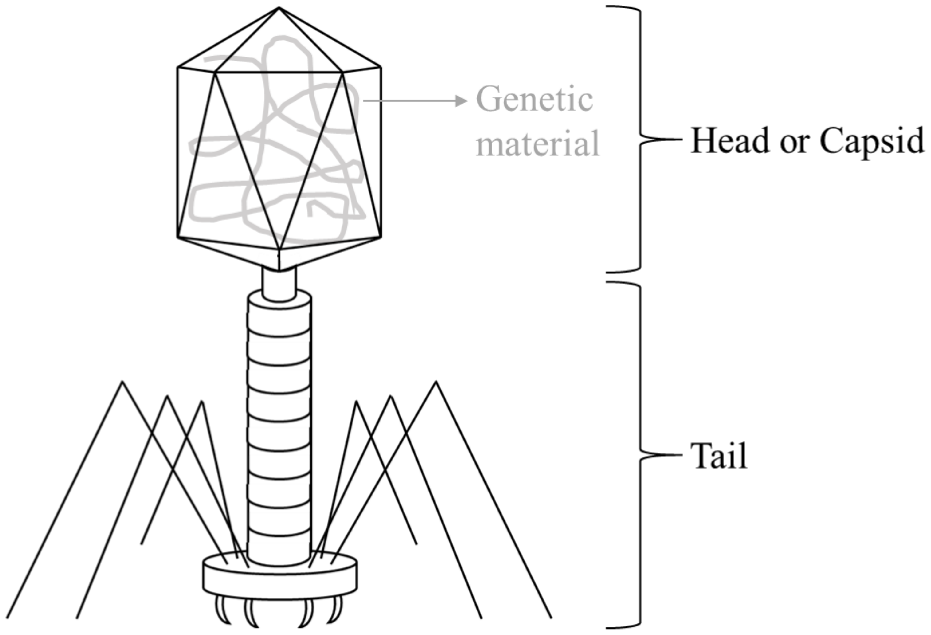

Figure 1. Diagram of a typical tailed phage with main components labelled.

Diversity

There are an estimated 10^{31} phage particles on Earth, making phages the most numerous entities in the world [3]. Phages are also thought to be the most diverse entity in the world and exist in every ecosystem [2]. Phage are able to survive not only in fresh water, seawater, forest floors, and soil, but also in extreme environments such as hot springs, polar inland waters, and the Sahara [2]. Surface seawater contains an estimated 10 million phages per milliliter [4] while soil contains an estimated 10^8 phages per gram [5]. With such astounding numbers, phages play significant roles in global biochemical cycles, bacterial host populations, and ecosystem functions [2].

The sequencing of all the DNA present in a sample is now possible thanks to advances in sequencing technology. The sequencing and analysis of all genetic material from an environmental sample is known as metagenomics [6]. Metagenomics can be implemented to study phages, such as identification and comparison of phages in a particular environment [6]. Importantly, metagenomics enables sequencing of organisms that are uncultivable in a laboratory setting. Approximately 95% of all bacteria cannot be cultivated in laboratory settings, so the phages that infect these bacteria also cannot be cultivated [6]. Such entities can now be studied via metagenomics. With genomic and metagenomics phage studies becoming increasingly feasible to conduct, novel information has been gathered on the genetics and environments of phages. Already, various metagenomic analyses have been conducted on water samples to study the phage diversity present. Metagenomic analysis of the freshwaters of the Amadorio River enabled the assembly of eight new, complete phage genomes [7]. Metagenomic analysis of 24 samples from various depths of the Mediterranean Sea enabled the assembly of 36 new, complete viral genomes, with many novel phages discovered, particularly from the depths [8].

Metagenomic analysis of the hypersaline Great Salt Lake enabled the identification of bacterial and phage communities present and gave insight into how phage-host interactions could be influencing the diversity, structure, and biogeochemical cycles of the lake [9]. Metagenomics has also granted novel insight into human phage communities. A great abundance and diversity of phages has been demonstrated in many of the human microbiomes, including the lung, vaginal, oral, intestinal, skin, and fecal microbiomes [10]. Of particular interest are phages present in the human gut microbiome, of which many studies have already been conducted. The composition and dynamics of the viral communities present in the human microbiome [11] along with the impact of phages on microbial activity and how this impacts gut homeostasis and human health [12] are being characterized on a large scale for the first time. However, metagenomics studies of phages are still novel and optimization of protocols for extraction of phages for metagenomics analysis is ongoing [13] [14].

Despite recent strides, there is still much unknown about the diversity of phages. To date only ~15,000 phages have been isolated and catalogued in the Actinobacteriophage Database [15]. The Actinobacteriophage Database is a comprehensive database containing information on the discovery, characterization, and genomics of phages that infect Actinobacteria [16]. Of those 15,000 catalogued phages, only ~3,000 have been sequenced [15]. Estimates are that less than 0.0002% of the global phage metagenome has been observed [17]. Even with such a small sample, a great diversity in phages has been observed. Sequenced mycobacteriophages have been categorized into clusters via their genomic similarity. A cluster is a group of phages containing at least 50% genomic similarity to each other. There are 122 clusters for the ~3,000 sequenced phages and 67 singleton phages (unique phages that do not fit into any existing cluster) [15]. Only the tip of the iceberg has been explored, and discovering novel phages from a variety of

environments and hosts will further expand our knowledge. Such discoveries will aid in the development and advancement of phage applications.

Applications of phages

A wide variety of applications using phages have been developed. Many applications are dedicated to improving human health, including phage therapy and the use of phages in dentistry. Other applications are dedicated to preventing disease in humans, such as the use of phages in food safety. Many other applications will be touched upon.

Antibiotic resistance is one of the greatest threats global health faces today. Humanity's excessive use of antibiotics to prevent and treat bacterial infections has promoted bacteria's development of mechanisms to resist antibiotics [18]. Antibiotics continue to decrease in effectiveness as antibiotic-resistant bacteria continue to increase [19]. The World Health Organization has stated that there must be an investment in research to develop new alternatives to antibiotics [20]. One such alternative is phage therapy. About a century ago, researchers studied the potential for phages as therapeutic agents to treat bacterial diseases. However, in the United States the rise of antibiotics led to a redirection in research interests. Now the field is reinvigorated, with knowledge of phages and development of phage therapy progressing every day. In 2019, two clinical trials involving phage-based drugs are beginning. The first is a phase I/II clinical trial to assess the safety and efficacy of EcoActive (a phage cocktail able to infect *Escherichia coli*) on intestinal adherent invasive *Escherichia coli* in patients with inactive Crohn's disease [21]. The second is a phase I/II clinical trial to assess the safety, tolerability, and efficacy of AB-SA01 phage therapy in combination antibiotic therapies in patients with ventricular assist devices infected by resistant *Staphylococcus aureus* [22]. The AB-SA01

clinical trial was fueled by the success of a phage cocktail that was administered in a patient suffering from a disseminated, resistant *Acinetobacter baumanni* pancreatic infection in an FDA-approved emergency investigational new drug application; the patient tolerated the therapy well and was able to make a full recovery [20].

There are many other medical applications for phages, such as use in dentistry. Dental plaque is a biofilm consisting mainly of bacteria, but also including fungi, protozoa, and viruses, with phages being the most common virus present [23]. Dental plaque is a precursor of many infectious diseases that impact oral health, including gingivitis, endodontic infections, peri-implantitis, and periodontal disease [24]. Current therapies for treating biofilm-derived infections are not specific, killing both protective and pathogenic bacteria [24]. The development of phage-based treatments to not only prevent, control, and treat oral infections, but also to potentially control the entire oral microbiome, are being investigated [24]. One example of such a treatment is the use of the phage strain phiIB-PAA2 to significantly reduce the population of the pathogenic *Pseudomonas aeruginosa* in biofilm starting just 2 hours after treatment [25].

Another important application of phages is their use in food safety. In the United States, approximately 48 million illnesses, 1,300 outbreaks, and 20 deaths occur each year due to foodborne diseases [26]. From 1998-2008, 45% of outbreaks with a known etiology were caused by bacteria, with *Salmonella* being one of the most common bacteria responsible [26]. Promising research has been conducted with the goal of determining if phages that infect *Salmonella* can reduce the levels of *Salmonella* present in various food products. When a pig model was challenged with *Salmonella typhimurium*, treatment with a phage cocktail reduced levels of *Salmonella* shedding in feces while still maintaining normal fecal flora [27]. When a chicken model was challenged with *Salmonella pullorum*, treatment with phage strain YSP2 significantly

reduced diarrhea and hemorrhaging of the intestine and liver [28]. Treatment of milk with phage strain P22 significantly decreased the growth of *Salmonella typhimurium* in the milk [29]. The application of phages to food safety is not limited to *Salmonella*. When a spinach model was challenged with *Listeria monocytogenes*, treatment with both a phage cocktail and modified atmosphere packaging significantly reduced the level of *Listeria* present [30]. Treatment of dairy cows with a phage cocktail demonstrated that all the *Staphylococcus aureus* strains found in dairy cows were susceptible to the phages [31]. Additionally, the development of very accurate reporter phage systems, which can rapidly detect viable pathogens in food, can be used as a biocontrol agent to monitor pathogen levels of foods [32].

While the use of phages in clinical, dental, and food safety applications is important, there are many other applications for phages being investigated. Other clinical applications of phages include vaccine development and delivery [33], treating open septic wounds and burn injuries [2], and maintaining a healthy microbiome [34]. Phages are also being applied to the sanitation of surfaces [35] and the treatment of wastewater [2]. The broad applications and high potential of phages demonstrates the importance of further studying phage biology and diversity. Continuing such studies will not only improve current applications, but potentially lead to new applications and new knowledge of phages.

Chapter 2: The role of enrichment in phage diversity

Introduction

Phages are plated with a species of host bacteria to give the phages a host to infect and reproduce. A common host bacteria is *Mycobacterium smegmatis* due to feasibility to use in a laboratory setting. Phages are isolated from environmental samples via two main techniques: direct plating and enrichment. Direct plating consists of attempting to yield phages directly from a sample without significantly altering the population present. In direct plating, phage buffer is added to the sample and the supernatant (which contains phages) is then plated with host bacteria. Enrichment consists of adding supplement, broth, inorganics, and host bacteria to the sample, incubating, and then plating the supernatant with host bacteria. By incubating the phages with host bacteria, phages are able to replicate greatly. The overall concentration of phages is increased, so enrichment is more likely to isolate phages compared to direct plating. Due to this advantage, performing enrichment is popular, with 77% of phages entries on The Actinobacteriophage Database indicating enrichment was utilized during isolation of phages.

Despite the great amount of diversity among phages, 20% of the ~3,000 sequenced phages on The Actinobacteriophage Database belong to just 1 of the 122 clusters. Additionally, of the phages found to infect *Mycobacterium smegmatis*, 33% belong to just 1 of the 29 clusters of phages that infect *Mycobacterium smegmatis* [15]. With such diversity present among phages, it is intriguing to observe relatively large proportions in only a few clusters, with the remaining clusters containing relatively few members. We wondered whether the observed distribution of phages in clusters actually reflects what is present in nature, or if instead there is some sort of skewing during the isolation process that makes it more likely to isolate a phage of a certain

cluster. As a significant procedure occurring between sample collection and isolation of phages from plaques on plates, we began to look into the enrichment procedure as a possible reason for the skewing of isolated phages.

We hypothesized that the enrichment procedure promotes the replication of phages belonging to only a few clusters, thus decreasing the diversity of phages that could be yielded from a sample. If true, the enrichment procedure could be significantly slowing the rate at which phages of rarer clusters are found, therefore decreasing the overall diversity of isolated phages. We would also like to gain novel insight into soil phage communities and implement nanopore sequencing to enable metagenomics analysis of such communities.

Results

Introduction to results

To conduct metagenomics analysis on soil phage communities, viral DNA was isolated from soil samples and sequenced. To determine how enrichment affects phage diversity, viral DNA was isolated from enriched soil samples and sequenced. The sequences were matched to reference phage genomes and the diversity of the samples was analyzed.

Sample collection and DNA isolation

Fifteen unique soil samples were collected. Viral DNA was isolated from each sample using the AllPrep PowerViral DNA Kit. Samples 1, 8, and 10 had enough DNA (>1 μg) for nanopore sequencing.

Samples 1, 8, and 10 were enriched for 48 hours in order to obtain data on how a population may be altered during enrichment. Aliquots were collected from the enrichment every 6 hours. Viral DNA was isolated from each enrichment aliquot using the AllPrep PowerViral DNA Kit. Across all samples, the enrichments at 24 and 48 hours had enough DNA (>1 μg) for nanopore sequencing.

DNA Detection

The concentration and quality of DNA isolated from the 15 unenriched soil samples were assessed using agarose gel electrophoresis. The resulting gel demonstrated relatively strong bands in samples 1, 2, 8, and 10 (Fig. 2). Following agarose gels, Qubit was used to accurately quantify the concentration of DNA present in each sample. A Qubit fluorometer uses fluorescent dye to determine the concentration of DNA in a sample via fluorochrome reactions with the

DNA [36]. The Qubit results showed that only samples 1, 8, and 10 had enough DNA (>1 μg) for nanopore sequencing.

DNA isolated from the enrichments of samples 1, 8, and 10 was analyzed via agarose gel electrophoresis to determine the relative amounts of DNA present. Across all 3 samples, the resulting gels demonstrated relatively strong bands at the 24- and 48-hour timepoints (Fig. 3). Due to having the strongest bands present in the agarose gel, the 24- and 48-hour timepoints for each of the 3 enrichments were subjected to Qubit concentration measurements. The Qubit results showed that, across all 3 samples, the 24- and 48-hour timepoints had enough DNA (>1 μg) for nanopore sequencing.

Immediately before sequencing, samples 1 and 10 were analyzed via TapeStation to confirm the quality and concentration of DNA a final time. The TapeStation system is a machine that carries out the electrophoresis of DNA libraries prepared for sequencing, determining the lengths and quantities of the DNA present [37]. Almost all TapeStation results showed bands approximately matching the original banding patterns of the agarose gels. The 48-hour timepoint for enriched sample 1 showed signs of DNA degradation and so was not sequenced. If possible, we would like to complete sequencing to have a larger sample size and characterize the population of phage present in sample 8.

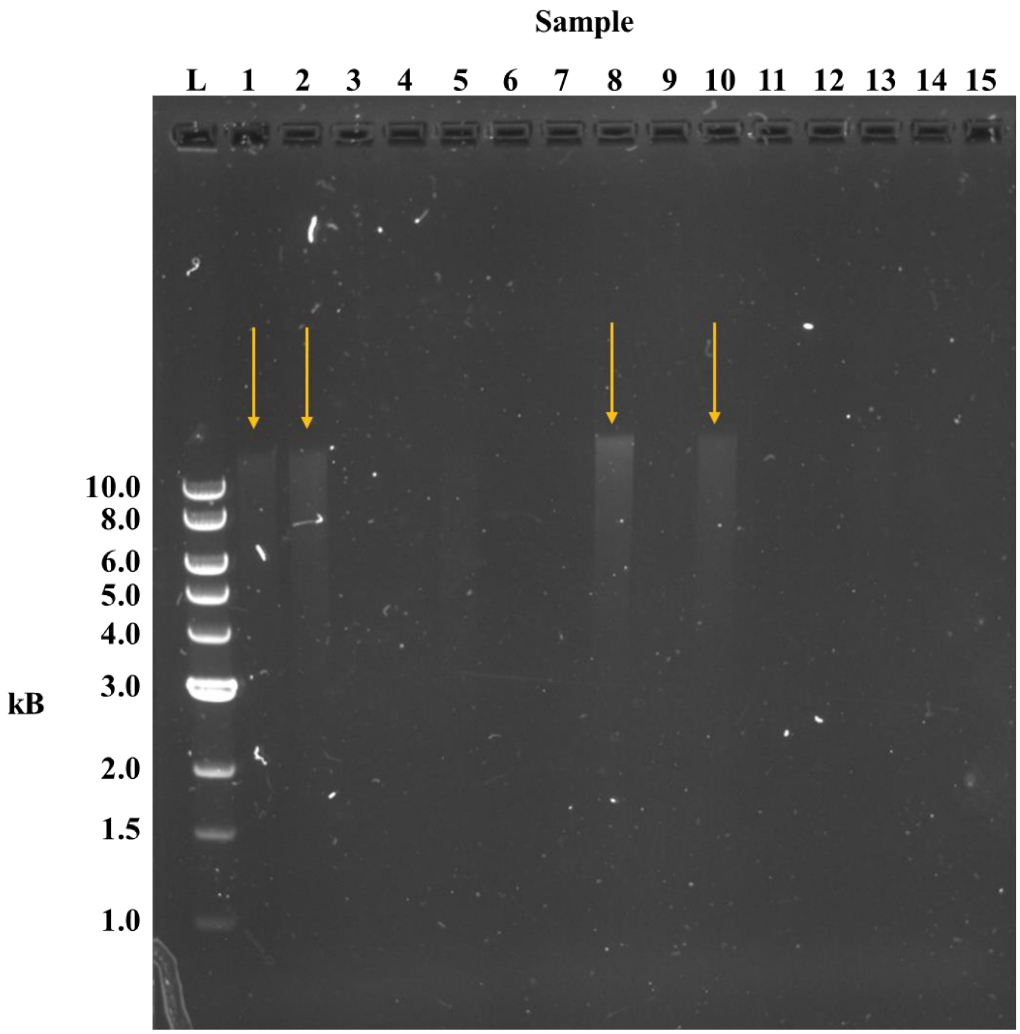

Figure 2. Agarose gel visualizing the DNA isolated from 15 soil samples. Each yellow arrow indicates the presence of DNA.

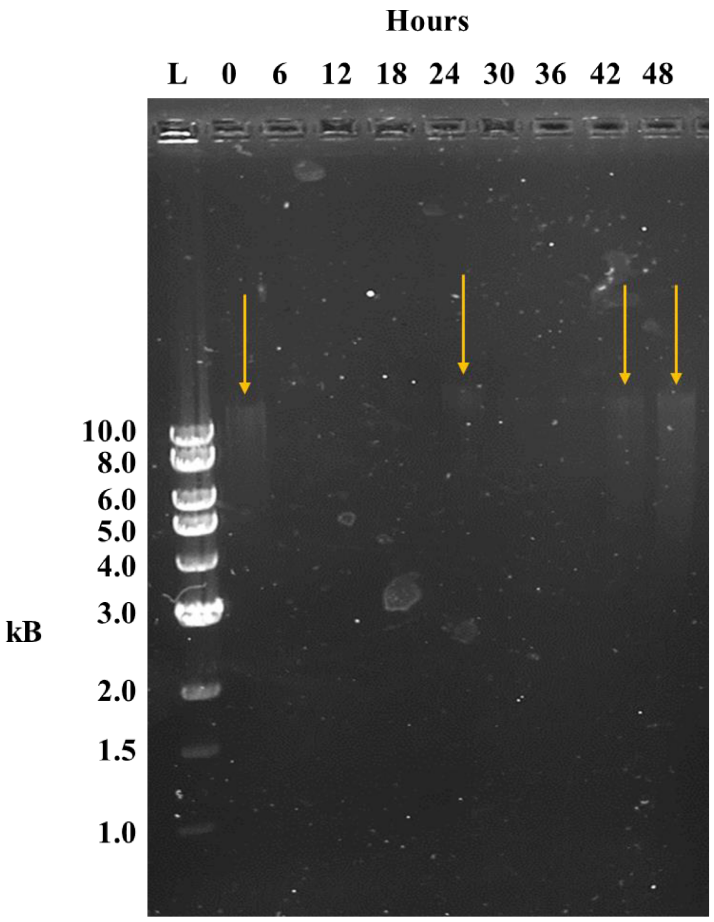

Figure 3. Representative agarose gel visualizing the DNA isolated the enrichments for sample 10. Each yellow arrow indicates the presence of DNA.

Results of nanopore sequencing

DNA samples were sequenced using nanopore methodology. Information regarding the number of sequencing reads, read lengths, and base-call quality was attained (Fig. 4). N50 is the value at which half of the reads are a length greater than the value. Overall, the N50 across all samples was ~9 kb, which is similar to the DNA lengths obtained from the agarose gels and TapeStation (Fig. 4A). The Phred algorithm is used in high throughput sequencing to transform the values of sequence features to a probability [38]. Phred quality scores were utilized to assess the quality of base-calling. For nanopore sequencing, base-calling is the process by which the electric signals are translated to nucleotides. Overall, the sequencing reads across all samples were within a range of ~8 to 12 for the base-call qualities, indicating ~84% to 94% accuracy in base-calling (Fig. 4B).

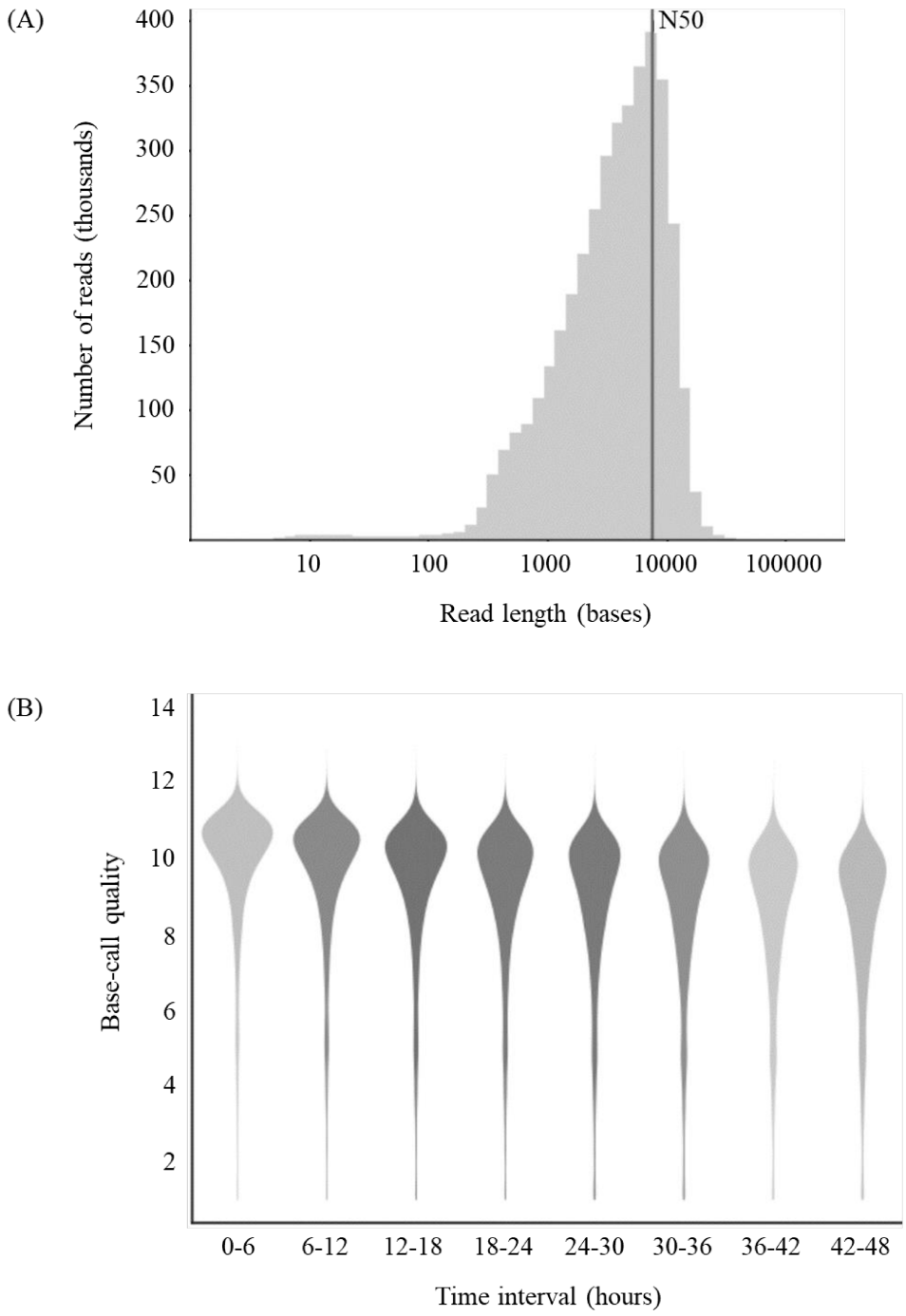

Figure 4. Representative sequencing run information. (A) Representative graph from enriched sample 1 at 24 hours which indicates the number of reads yielded for each read length. (B) Representative graph from unenriched sample 1 which indicates the base call quality over the 48-hour sequencing period.

Sequence matches to known phages

The sequences were compared to known phages in the Actinobacteriophage Database by the sequencing read alignment program Minimap2. Minimap2 is a general-purpose alignment program used to map long DNA sequences against a reference database, and has a higher accuracy for alignment of long DNA sequences compared to other mainstream long-read mappers [39]. In order to be matched, a sequence had to be at least 85% identical in sequence to the known phages. If a sequence was matched to multiple reference genomes, only the best match was counted. The proportion of sequences that matched known phages was identified (Fig. 5). The unenriched samples 1 and 10 each have ~1% of the sequences matched to known phages (Fig. 5). Of intrigue is the significantly lower percentage of reads matched to known phages in the 48 hour timepoint of sample 10 (Fig. 5). While the reason for the lower percentage of matches is unknown, it is of interest to further study this timepoint in the future to discern why this occurred.

Bacterial hosts of the sequences

The sequences matched to known phages were compiled. To observe the distribution of phage types present, the bacterial host infected by each matched sequence was gathered (Fig. 6). Of the phages identified, a relatively high proportion infecting *Mycobacterium* were represented. There are also relatively high proportions of phages infecting *Gordonia*, *Microbacterium*, and *Streptomyces* present. When analysis of bacterial hosts was conducted on the enriched samples, all enriched samples consisted of at least 99% phages infecting *Mycobacterium*.

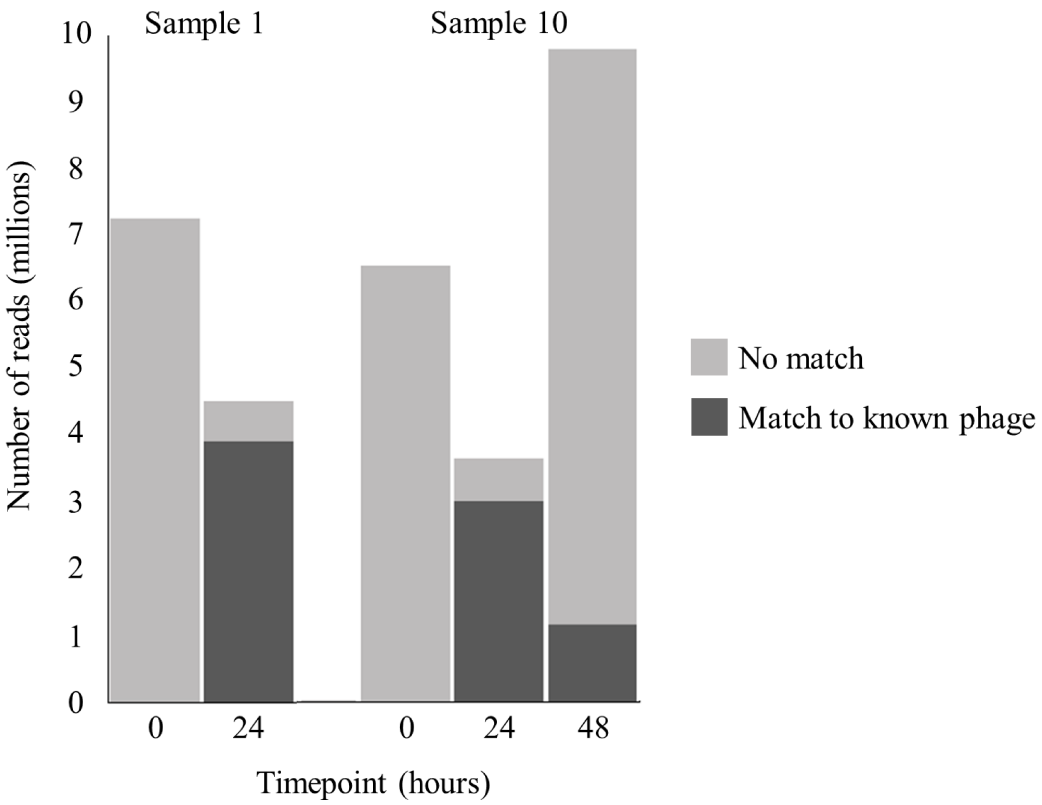

Figure 5. Resulting reads of the sequencing data, with the proportion of sequences that are unmatched or matched to known phage sequences.

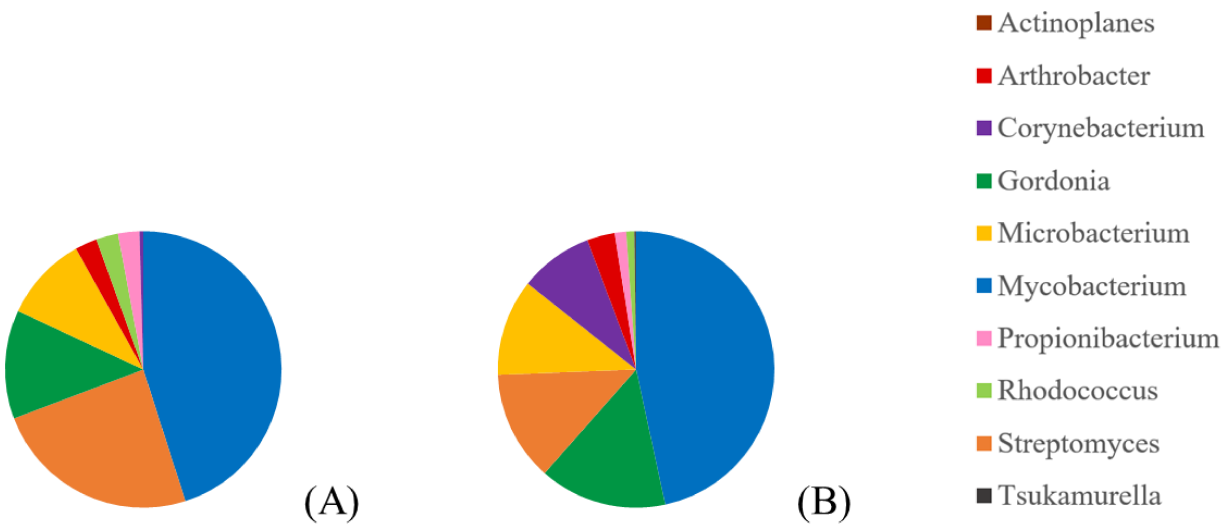

Figure 6. The bacterial hosts of the sequences matched to known phages. (A) Unenriched sample 1. (B) Unenriched sample 10.

Clusters of the sequences

To evaluate the effect of enrichment on the diversity of the phage community in each sample, the clusters of the sequences were identified. The sequences matched to known *Mycobacterium smegmatis* phages were compiled. Since *Mycobacterium smegmatis* was used as the host bacteria for enrichment, only the *Mycobacterium smegmatis* phages in the unenriched soil samples were taken into account for analysis on population trends pre and post enrichment. Additionally, for the enriched samples there are no significant numbers of sequences that match to known phages that infect other hosts. Information on the clusters of the sequences (based on their known phages) was attained. To observe the diversity of the sequences present, the clusters of the sequences matched to known phages were identified (Fig. 7).

Although relatively high percentages of sequences that matched to a known phage sequence belonged to a limited number of clusters, such as C, A, and F, approximately 20 clusters are present in each unenriched sample (Fig. 7A and 7C). Of important notice in the unenriched samples is the high proportion of matches to singletons, indicating there are sequences present that could cluster to known singletons.

In comparison, the enrichments have less diversity at the cluster level, with each enrichment having over half of the matched sequences belong to a single cluster (Fig. 7B, 7D, and 7E). Interestingly, the enrichments for sample 1 have a large proportion of matches to cluster A while the enrichments for sample 10 have a large proportion of matches to clusters A and B. The A and B clusters are currently the two most numerous clusters for *Mycobacterium smegmatis* phages isolated via enrichment. Therefore, the data supports the idea that enrichment could be causing a higher proportion of A and B cluster phages to be yielded in laboratory settings via enrichment than there actually are in nature.

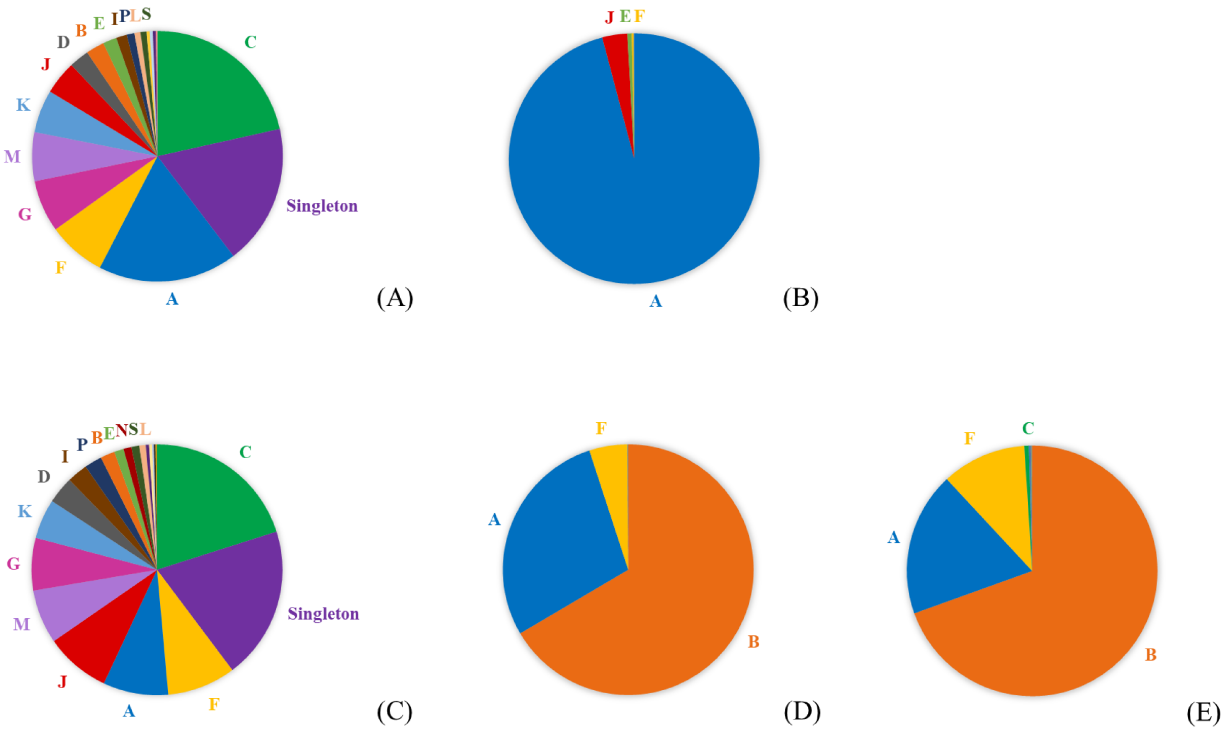

Figure 7. The clusters of the sequences matched to known phages able to infect *Mycobacterium smegmatis*. (A) Unenriched sample 1. (B) 24-hour enrichment, sample 1. (C) Unenriched sample 10. (D) 24-hour enrichment, sample 10. (E) 48-hour enrichment, sample 10.

Subclusters of the sequences

After observing skew in diversity at the cluster level, we wanted to see if skew continued at the subcluster level. Once again, the sequences matched to known *Mycobacterium smegmatis* phages were compiled and only the *Mycobacterium smegmatis* phages in the unenriched soil samples should be taken into account for analysis on population trends pre and post enrichment. For each sample, the cluster exhibiting the greatest change due to enrichment was identified: for sample 1 this is cluster A and for sample 10 this is cluster B. Information on the subclusters of the sequences (based on their known phages) was attained. To observe the diversity of the sequences present within the cluster exhibiting the greatest change, the subclusters of the sequences matched known phages was identified (Fig. 8).

The unenriched sample 1 exhibits a moderate amount of diversity, with prominent amounts of five of the twenty subclusters within the A cluster (Fig 8A). After 24 hours of enrichment, sample 1 contains predominantly phages of the A1 subcluster, as well as A11 subcluster phages (Fig. 8B). Although not displayed, one piece of important data is the change undergone by C cluster phages due to enrichment of sample 1. The C cluster consists of two subclusters: the much more numerous C1 subcluster, and the rarer C2 subcluster. In the unenriched sample 1, the vast majority of C cluster phages belong to the C2 subcluster. Post-enrichment, the vast majority of C cluster phages now belong to the C1 subcluster, supporting the idea that enrichment is skewing phage diversity. This could be a possible explanation as to why the C1 subcluster contains 132 members while the C2 subcluster only contains 2 members [15].

The unenriched sample 10 also exhibits a moderate amount of diversity, with prominent amounts of four of the nine subclusters within the B cluster (Fig 8C). After both 24 and 48 hours

of enrichment, sample 10 contains predominantly phages of the B1 subcluster (Fig. 8D and 8E). Intriguingly, once again the same trend is observed in the C phage subclusters: pre-enrichment the vast majority of C cluster phages belong to the C2 subcluster while post-enrichment, the vast majority of C cluster phages now belong to the C1 subcluster.

As with the analysis of the clusters, analysis of the subclusters demonstrates a decrease in diversity as a result of enrichment. Intriguingly, the A1 and B1 subclusters are the largest subclusters within their respective clusters, further supporting the idea that enrichment could be skewing the types of phages yielded in laboratory settings than are actually in nature. The trends observed in the C1 and C2 subclusters are further support of skewing. However, it is important to note that there are exceptions. For example, in both samples 1 and 10, the number of F1 subcluster phages is larger than F2 subcluster phages both pre and post enrichment. Perhaps the skewing enrichment likely induces affects some clusters and subclusters more than others.

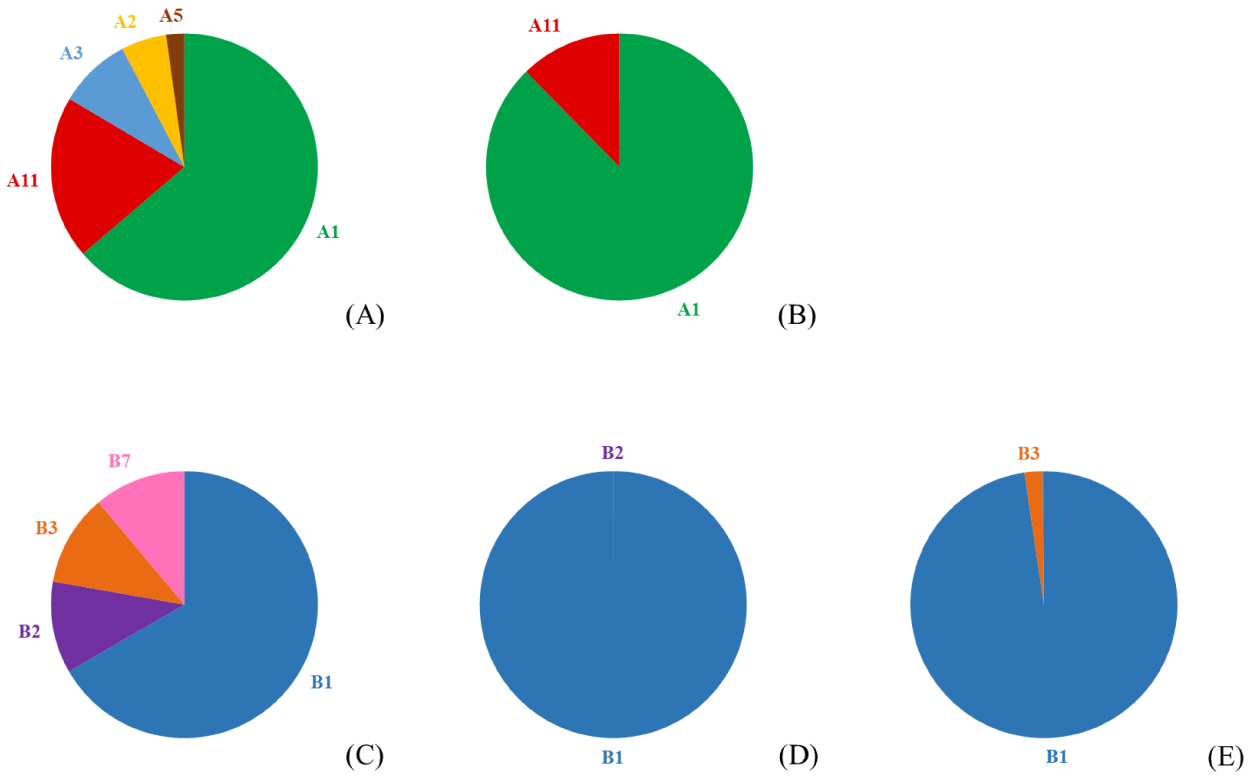

Figure 8. The subclusters of the overall most prominent cluster in each sample for the sequences matched to known phages able to infect *Mycobacterium smegmatis*. (A) Unenriched sample 1. (B) 24-hour enrichment, sample 1. (C) Unenriched sample 10. (D) 24-hour enrichment, sample 10. (E) 48-hour enrichment, sample 10.

Trends of the sequences

After observing skew in diversity at both the cluster and subcluster levels, we wanted to see if skew continued at the individual phage level. The trends of individual reference phages from the cluster exhibiting the greatest change due to enrichment were plotted (Fig. 9). To account for the differences in the total number of sequences produced by each sample's sequencing run, the percentage of sequences matched to the phage was used instead of the raw number of sequences. Overall, in both samples there is no clear dominance of a few individual phages, though some phages have greater increases than others. In sample 1, those reference phages exhibiting the greatest increases are of the subcluster A1 (Fig 9A). In sample 10, those reference phages exhibiting the greatest increases are of the subcluster B1 (Fig 9B). This supports the previous findings that there is skewing present at the subcluster level. However, there is no support for skewing at the individual phage level. This could be an indication that phages of the same subcluster contain very similar genetics that enable skew of the subcluster to occur.

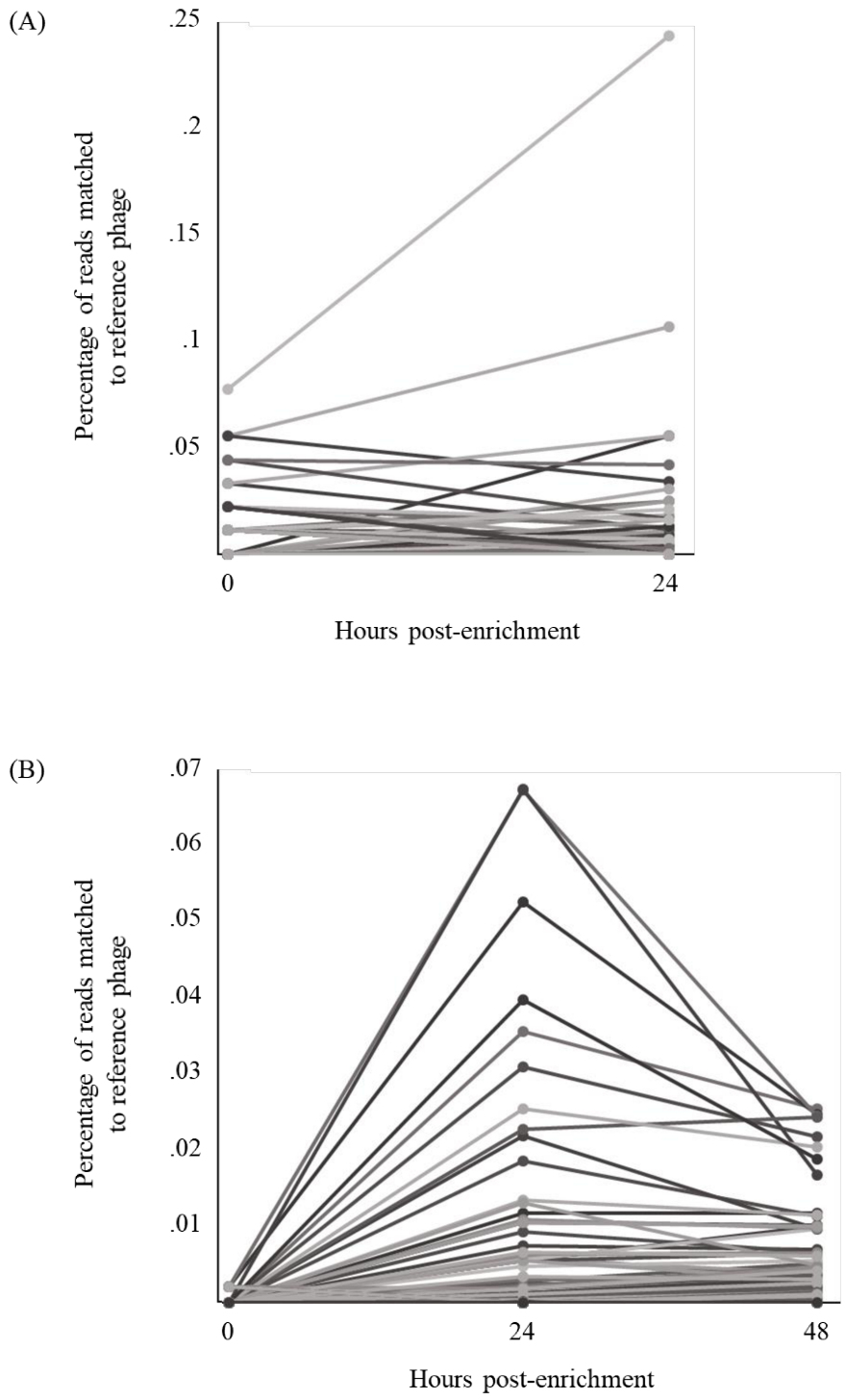

Figure 9. The trends of individual identified phages of the overall most prominent cluster in each sample over the time. The percentage of sequences matched to each identified phage at each timepoint is displayed. (A) A cluster phages of sample 1. (B) B cluster phages of sample 10.

Discussion

The purpose of enrichment is to increase the numbers of phages able to infect the bacteria included in the enrichment culture. The metagenomic analysis demonstrates that enrichment does achieve its purpose, however enrichment could come with consequences. The high percentage of phages infecting *Mycobacterium* in the enriched samples indicates that enrichment using *Mycobacterium smegmatis* as the host does indeed achieve its purpose of amplifying the number and proportion of phages infecting *Mycobacterium* originating from a soil sample. Overall, the sequencing data demonstrates a decrease in diversity of phages in a sample post-enrichment. Additionally, the most prominent clusters present in the enriched samples are the largest known clusters, indicating that enrichment with *Mycobacterium smegmatis* is likely promoting the growth of A and B cluster phages more so than others. Contrasting results such as the maintenance of the F cluster and proportions of the F1 and F2 subclusters in sample 10 suggest that enrichment does not affect all phages equally, with the promotion or hindrances of some clusters and other clusters unaffected. Additionally, the novel workflow and sequencing analysis appears successful in attaining metagenomics information for both unenriched and enriched samples.

However, it is important to note that this study is currently limited by the number of known phage genomes. With such a small percentage of phages having been sequenced, fewer sequences are able to be matched to reference genomes, leaving more of the sequences unidentified. Further sequencing and characterization of phages would yield more robust data that could shed more light on exactly how enrichment is affecting phages of each cluster. Another limitation is the lack of characterization of reproductive aspects of phages, such as the latent period, eclipse phase, burst size, and more. Experiments seeking to characterize such

aspects of phage replication would clarify how replication differs between clusters. With this information, one might be able to discern if these additional factors are also influencing the numbers of phages resulting from enrichment and could be responsible for the skewing of diversity that has been observed. In particular, it would be intriguing to study the replication of A and B cluster phages to seek explanation for the observed increase during enrichment. Additionally, in the future increasing the sample size of this study by analyzing additional soil samples using the same methods could produce further corroboration for the results of this study.

With advancements in sequencing technology, more information has been able to be discerned about phage genomics and metagenomics than ever before. However, with only 0.0002% of the global phage metagenome observed and there is still much to discover [17]. This study aimed to determine if the enrichment procedure is causing a decrease in the diversity of phages yielded during isolation, and the data supports the hypothesis. With the growing potential for phage applications, it is important to continue to isolate and characterize phages.

Methods

Collection of samples

Soil samples were collected from random locations across the Johns Hopkins University Homewood campus, with locations sufficiently isolated from each other. A 5 cm by 5 cm plot 2.5 cm deep was made and enough soil was collected from the plot to fill a 50 mL conical tube. Processing of samples began immediately. Sample 1 was collected from Wyman Quad (DMS coordinates: 39°19'40" N 76°37'13" W), sample 8 from near the Recreational Center (DMS coordinates: 39°19'55" N 76°37'13" W), and sample 10 from near Stony Run (DMS coordinates: 39°19'55" N 76°37'24" W).

Enrichment of samples

Enrichment was performed according to SEA-Phages Laboratory Manual protocol, except for the following deviations: 2 g of soil was added to the enrichment flask; enrichment was carried out for a total of 48 hours. Starting once incubation begins, a 2 mL aliquot was collected from the flask every 6 hours. The *Mycobacterium smegmatis* colony used in enrichment was started from a frozen stock of MC^2155 strain *Mycobacterium smegmatis*. 7H9/glycerol broth was made by our lab using Difco Middlebrook 7H9 Broth (Becton, Dickinson, and Company; lot number 2003737) and 99%+ glycerol (Alfa Aesar, C29Y030). BBL Middlbrook ADC Enrichment (Becton, Dickinson, and Company; lot number 8131952) was purchased for enrichment.

Isolation of DNA

The AllPrep® PowerViral® DNA/RNA Kit (Qiagen, lot number 160018997) was used to isolate phage DNA from the soil samples and enriched samples. The kit was used according to the manufacturer's protocol. The beginning of the procedure differed for the soil and enriched samples. To break up the soil and more effectively isolate phage DNA, the soil samples were subjected to bead beating (steps 3-7 of the protocol). The enriched samples did not need beadbeating and therefore steps 3-7 were skipped.

Agarose gel electrophoresis

1% agarose solution was created using agarose (Sigma-Aldrich, lot number SLBD2493V) and 1X TAE. 1 µL SYBR Safe DNA gel stain (Invitrogen by Thermo Fisher Scientific, lot number 1876635) was added for every 10 mL of 1% agarose solution. Bio-Rad gel box (model Mini-Sub Cell GT) was set up and used according to manufacturer's protocol. For the ladder, 1 µL of 1 kb DNA Ladder #N3232S (New England Biolabs, lot number 0730810) was combined with 5 µL of 6X Purple Loading Dye (New England Biolabs, lot number 0201703) that has been diluted to 1X with UltraPure water. For each DNA sample, 1 µL of sample was combined with 5 µL of the diluted 1X Purple Loading Dye. Gel was run at 90 V for 45 minutes. Gel was imaged using a ProteinSimple Imager (model FluorChem M) on the "Ethidium Bromide" setting.

TapeStation

Preparation of samples, loading of Agilent 4200 TapeStation, and programming of TapeStation was done according to the manufacturer's protocol. Genomic DNA ScreenTape was

utilized. The Genomic DNA Ladder was the ladder used (Agilent, lot number 0006435306). The reagent used in the preparation of samples was the Genomic DNA Sample Buffer (Agilent, lot number 0006435306).

Qubit

Preparation of samples and programming of Invitrogen Qubit 3 Fluorometer was done according to the manufacturer's protocol. The reagents used were from the Qubit dsDNA HS Assay Kit (Invitrogen, lot number 2016949). The Qubit dsDNA HS Standard #1 and Qubit dsDNA HS Standard #2 were used as standards.

Nanopore sequencing

Oxford Nanopore sequencing technology was used to sequence the samples. Preparation of samples was done according to the manufacturer's protocol (protocol "1D Genomic DNA by Ligation (SQK-LSK109)"), version GDE_9063_v109_revD_23May2018), except for the following alterations: skip the optional DNA fragmentation steps; for the DNA repair and end-prep steps: 50 µL DNA sample, 7 µL Ultra II End Prep Reaction Buffer #E7647AA (New England Biolabs, lot number 0111801), and 3 µL Ultra II End Prep Enzyme Mix #E7647AA (New England Biolabs, lot number 0061801) were combined in a PCR tube and then incubated at 20° C for 20 minutes and 65° C for 20 minutes; for AMPure XP bead clean-up steps, 60 µL of resuspended AMPure XP beads were added for every 1X of pooled sample volume. The Ligation Sequencing Kit (Oxford Nanopore Technologies, SQK-LSK109) was used for adapter ligation and clean-up. Each of the reagents in the kit come from the following batches: Ligation Buffer (LNB) is from batch SK1461004, Adapter Mix (AMX) is from batch SK1451005, Long Fragment Buffer (LFB) is from batch SK1471002, Elution Buffer (EB) is from batch

SK1491004, Flush Buffer (FLB) is from batch SK1291014, Flush Tether (FLT) is from batch SK1301012, Loading Beads (LB) is from batch SK1271008, and Sequencing Buffer (SQB) is from batch SK1282006. NEBNext Quick T4 DNA Ligase (New England Biolabs, lot number 1231803) was also used during adapter ligation and clean-up steps.

Each prepped sample (containing ≤1 ug of DNA) was loaded into a R9.4.1 FLO-MIN106 flow cell. Each flow cell was loaded into a GridION X5 system. Programming of 48-hour sequencing run was done according to manufacturer's protocol (protocol "1D Genomic DNA by Ligation (SQK-LSK109)"). Execution of sample preparation and nanopore sequencing was done in collaboration with the Timp Lab (Johns Hopkins University, Department of Biomedical Engineering).

Matching of sequencing runs to known phages

The sequences that resulted from nanopore sequencing were matched to known phage genomes via the alignment program Minimap2. Alignment was carried out according to program's protocol. A library of reference genome was constructed for use in Minimap2 and consists of all Actinobacteriophage genomes published in the Actinobacteriophages Database as of 2018. Execution of sequencing read alignment was done in collaboration with the Timp Lab (Johns Hopkins University, Department of Biomedical Engineering).

References

[1] K. Chamakura and R. Young, "Phage single-gene lysis: Finding the weak spot in the bacterial cell wall," *Journal of Biological Chemistry,* vol. 294, no. 10, p. 3350–3358, 2019.

[2] S. Sharma et al., "Bacteriophages and its applications: an overview," *Folia Microbiologica,* vol. 62, pp. 17-55, 2017.

[3] M. L. Pedulla, "Origins of Highly Mosaic Mycobacteriophage Genomes," *Cell,* 2003.

[4] M. Breitbart, "Marine viruses: truth or dare.," *Annual Review of Marine Science. ,* vol. 4, p. 425–448, 2012.

[5] K. Williamson, K. Wommack and M. Radosevich, "Sampling natural viral communities from soil for culture-independent analyses.," *Applied and Environmental Microbiology,* vol. 69, no. 11, pp. 6628-33, 2003.

[6] M. R. Clokie, A. D. Millard, A. V. Letarov and S. Heaphy, "Phages in nature," *Bacteriophage,* 2011.

[7] R. Ghai et al., "Metagenomic recovery of phage genomes of uncultured freshwater actinobacteria," *The ISME Journal,* vol. 11, no. 1, pp. 304-308, 2017.

[8] M. Lopez-Perez et al., "Genome diversity of marine phages recovered from Mediterranean metagenomes: Size matters," *PLOS Genetics,* vol. 13, no. 9, 2017.

[9] A. M. Motlagh et al., "Insights of Phage-Host Interaction in Hypersaline Ecosystem through Metagenomics Analyses," *Frontiers in Microbiology,* vol. 8, 2017.

[10] F. Navarro and M. Muniesa, "Phages in the Human Body," *Frontiers in Microbiology,* vol. 8, 2017.

[11] V. Aggarwala et al., "Viral communities of the human gut: metagenomic analysis of composition and dynamics," *Mobile DNA,* vol. 8, no. 12, 2017.

[12] P. Lepage, "A metagenomic insight into our gut's microbiome.," *Gut,* vol. 62, no. 1, pp. 146-58, 2013.

[13] J. Castro-Mejía et al., "Optimizing protocols for extraction of bacteriophages prior to metagenomic analyses of phage communities in the human gut," *Microbiome,* vol. 64, no. 3, 2015.

[14] M. Muhammed et al., "Metagenomic Analysis of Dairy Bacteriophages: Extraction Method and Pilot Study on Whey Samples Derived from Using Undefined and Defined Mesophilic Starter Cultures," *Applied and Environmental Microbiology,* vol. 83, no. 19, 2017.

[15] "The Actinobacteriophage Database," [Online]. Available: https://phagesdb.org/. [Accessed March 2019].

[16] D. Russell and G. Hatfull, "PhagesDB: the actinobacteriophage database," *Bioinformatics,* vol. 33, no. 5, pp. 784-786, 2017.

[17] F. Rohwer, "Global Phage Diversity," *Cell,* 2003.

[18] L. Dever and D. TS, "Mechanisms of bacterial resistance to antibiotics.," *Archives of Internal Medicine,* vol. 151, no. 5, pp. 886-895, 1991.

[19] J. Martinez, "General principles of antibiotic resistance in bacteria," *Drug Discovery Today: Technologies,* vol. 11, pp. 33-39, 2014.

[20] R. Schooley and e. al, "Development and Use of Personalized Bacteriophage-Based Therapeutic Cocktails To Treat a Patient with a Disseminated Resistant Acinetobacter baumannii Infection," *Antimicrobial Agents and Chemotherapy,* vol. 61, no. 10, 2017.

[21] "Safety and Efficacy of EcoActive on Intestinal Adherent Invasive E. Coli in Patients With Inactive Crohn's Disease," U.S. National Library of Medicine, 4 March 2019. [Online]. Available: https://www.clinicaltrials.gov/ct2/show/NCT03808103?term=phage&draw=1&rank=13. [Accessed 7 April 2019].

[22] "Individual Patient Expanded Access for AB-SA01, an Investigational Anti-Staphylococcus Aureus Bacteriophage Therapeutic," U.S. National Library of Medicine, 10 January 2018. [Online]. Available: https://www.clinicaltrials.gov/ct2/show/NCT03395769?term=AB-SA01&rank=1. [Accessed 7 April 2019].

[23] G. Pinto et al., "The role of bacteriophages in periodontal health and disease," *Future Microbiology,* vol. 11, no. 10, 2016.

[24] M. Shlezinger et al., "Phage Therapy: A New Horizon in the Antibacterial Treatment of Oral Pathogens.," *Current Topics in Medicinal Chemistry,* vol. 17, no. 10, pp. 1199-1211, 2017.

[25] D. Pires et al., "Use of newly isolated phages for control of Pseudomonas aeruginosa PAO1 and ATCC 10145 biofilms," *Research in Microbiology,* vol. 162, pp. 798-806, 2011.

[26] L. Gould et al., "Surveillance for foodborne disease outbreaks - United States, 1998-2008.," *Morbidity and mortality weekly report. Surveillance summaries.,* vol. 62, no. 2, pp. 1-34, 2013.

[27] B.-J. Seo et al., "Evaluation of the broad-spectrum lytic capability of bacteriophage cocktails against various Salmonella serovars and their effects on weaned pigs infected

with Salmonella Typhimurium," *Journal of Veterinary Medical Science,* vol. 80, no. 6, pp. 851-860, 2018.

[28] K. Tie et al., "Isolation and identifcation of Salmonella pullorum bacteriophage YSP2 and its use as a therapy for chicken diarrhea," *Virus Genes,* vol. 54, pp. 446-456, 2018.

[29] W. Phongtang et al., "Bacteriophage control of Salmonella Typhimurium in milk," *Food Science and Biotechnology,* vol. 28, no. 1, pp. 297-301, 2019.

[30] O. Boyacioglu, A. Sulakvelidze, M. Sharma and I. Goktepe, "Effect of a bacteriophage cocktail in combination with modified atmosphere packaging in controlling Listeria monocytogenes on fresh-cut spinach.," *Irish Journal of Agricultural & Food Research,* vol. 55, no. 1, pp. 74-79, 2016.

[31] D. Varela-Ortiz et al., "Antibiotic susceptibility of Staphylococcus aureus isolated from subclinical bovine mastitis cases and in vitro efficacy of bacteriophage," *Veterinary Research Communications,* vol. 42, pp. 243-250, 2018.

[32] J. Bai et al., "Biocontrol and Rapid Detection of Food-Borne Pathogens Using Bacteriophages and Endolysins," *Frontiers in Mircobiology,* 2016.

[33] L. Rahbarniaa et al., "Evolution of phage display technology: from discovery to application," *Journal of Drug Targeting,* vol. 25, no. 3, pp. 216-224, 2017.

[34] L. R. Lopetuso, M. E. Giorgio, A. Saviano, F. Scaldaferri, A. Gasbarrini and G. Cammarota, "Bacteriocins and Bacteriophages: Therapeutic Weapons for Gastrointestinal Diseases?," *International Journal of Molecular Sciences,* 2019.

[35] M. D'Accolti et al., "Efficient removal of hospital pathogens from hard surfaces by a combined use of bacteriophages and probiotics: potential as sanitizing agents.," *Infection and Drug Resistance,* vol. 30, no. 11, pp. 1015-1026, 2018.

[36] G. Ponti and e. al, "The value of fluorimetry (Qubit) and spectrophotometry (NanoDrop) in the quantification of cell-free DNA (cfDNA) in malignant melanoma and prostate cancer patients," *Clinica Chimica Acta,* 2018.

[37] C. Hussing and e. al, "Quantifcation of massively parallel sequencing libraries – a comparative study of eight methods," *Scientific Reports,* 2018.

[38] S. Zhang and e. al, "Estimating Phred scores of Illumina base calls by logistic regression and sparse modeling.," *BMC Bioinformatics,* vol. 18, no. 1, 2017.

[39] H. Li, "Minimap2: pairwise alignment for nucleotide sequences.," *Bioinformatics,* vol. 34, no. 18, pp. 3094-3100, 2018.

Vanessa Gonzalez
9 East 33rd Street, Apt. 302B, Baltimore, MD 21218 | (908)-565-1129 | vgonza16@jhu.edu

EDUCATION

Princeton University Princeton, NJ
Ph.D. in Molecular Biology Beginning August 2019

Johns Hopkins University Baltimore, MD
M.S. in Molecular and Cellular Biology Expected May 2019
- Thesis: The role of enrichment in decreasing diversity of mycobacteriophage genome databases
- Relevant Coursework: Mentored Research, Advanced Seminar: Molecular and Cellular Biology

B.S. in Molecular and Cellular Biology, Additional Major in Spanish Expected May 2019
- Relevant Coursework**:** Phage Research, Cell Biology with lab, Genetics with lab, Developmental Biology, Biochemistry, Protein Engineering and Biochemistry Lab, Organic Chemistry with lab

RESEARCH EXPERIENCE

Dr. Joel Schildbach's Laboratory (JHU) Baltimore, MD
Lab Member August 2015-Present
- Investigate how enrichment affects the diversity of phage in soil samples
- Utilize metagenomics and nanopore sequencing to isolate DNA of phage populations from distinct soil samples
- Isolate and characterize novel bacteriophages
- Sequence and annotate bacteriophage genomes, identifying potential genes and their functions
- Acquired basic laboratory skills, including bacterial cell culture, agarose gel electrophoresis, etc.

HONORS AND AWARDS

General Honors Expected May 2019

Departmental Honors May 2019
- Expected for both the Molecular and Cellular Biology major and the Spanish major

Dean's List May 2016-Present
- Earned the following semesters: Spring 2016, Fall 2016, Fall 2017, Spring 2018, Fall 2018

PROFESSIONAL EXPERIENCE

Johns Hopkins Biology Department Baltimore, MD
Teaching Assistant August 2018-Present
- Teach and grade coursework for a section of 24 students for General Biology Lab
- Proctor and grade exams for General Biology lecture

Bristol-Myers Squibb Hopewell, NJ
Intern June 2018-August 2018
- Conducted a study, under the supervision of mentor, of the proteome of distinct T-cell subtypes to identify potential distinguishing biomarkers and IO targets
- Implemented techniques such as trypsin digest, column purification, and mass spectrometry
- Analyzed resulting mass spectrometry data to identify signature proteins
- Assisted in the study of the effect of an experimental fibrosis drug in mice

PILOT Learning Baltimore, MD
Tutor August 2017-May 2018

- Led study teams of 10 students each who are taking Calculus II for Biological Sciences
- Conducted weekly 2 hour session with each team to review topics and complete problem sets
- Aided students in improving their academic and studying skills and provided support

SYMPOSIUM AND PRESENTATION EXPERIENCE

Molecular and Cellular Biology Master's Program 2019 Thesis Presentations Expected May 2019

- Will give a 30 minute research presentation as part of the culmination of Master's degree

Bristol-Myer Squibb's Summer Intern Research Symposium August 2018

- Presented a research poster at the conclusion of research-based internship to a wide audience consisting of researchers, administrators, and other interns

VOLUNTEER WORK

MAPP (Mentoring Assistance Peer Program) Baltimore, MD
Underrepresented Freshmen Counselor and Mentor August 2016-Present

- Counsel and mentor three underrepresented freshmen by providing academic and personal development skills and support, as well as serve as liaison to university student support services
- Implement with a team of other mentors various academic, cultural, and service based enrichment events and programs for freshmen mentees throughout academic year

Tutorial Project Baltimore, MD
Tutor January 2016-May 2017

- Tutored elementary school student for 4 hours per week following agreed upon learning outcomes using various techniques and implementing several learning styles

TECHNICAL SKILLS

Wet Lab: Miniprep, PCR, agarose gel electrophoresis, Western blot, UV-Vis spectroscopy, mass spectrometry, column purification, bacterial cell culture techniques, etc.
Computer: Microsoft Excel, PowerPoint, and Word, PyMOL (Python), BLAST
Languages: Spanish